J Nonverbal Behav (2014) 38:31–52
DOI 10.1007/s10919-013-0165-x

ORIGINAL PAPER

The Role of Perceived Voice and Speech Characteristics in Vocal Emotion Communication

Tanja Bänziger · Sona Patel · Klaus R. Scherer

Published online: 17 October 2013

Abstract Aiming at a more comprehensive assessment of nonverbal vocal emotion communication, this article presents the development and validation of a new rating instrument for the assessment of perceived voice and speech features. In two studies, using two different sets of emotion portrayals by German and French actors, ratings of perceived voice and speech characteristics (loudness, pitch, intonation, sharpness, articulation, roughness, instability, and speech rate) were obtained from non-expert (untrained) listeners. In addition, standard acoustic parameters were extracted from the voice samples. Overall, highly similar patterns of results were found in both studies. Rater agreement (reliability) reached highly satisfactory levels for most features. Multiple discriminant analysis results reveal that both perceived vocal features and acoustic parameters allow a high degree of differentiation of the actor-portrayed emotions. Positive emotions can be classified with a higher hit rate on the basis of perceived vocal features, confirming suggestions in the literature that it is difficult to find acoustic valence indicators. The results show that the suggested scales (Geneva Voice Perception Scales) can be reliably

T. Bänziger conducted Study 1 as part of an unpublished doctoral dissertation (supervised by K. R. Scherer).

Preliminary results of Study 1 have been presented at an International Speech Communication Association (ISCA) satellite workshop in Geneva in 2003.

S. Patel conducted the second rating study, performed the extraction of acoustical parameters for Study 2, and presented a preliminary report of Study 2 at an Acoustical Society of America (ASA) meeting in Chicago in 2011.

T. Bänziger · S. Patel · K. R. Scherer (✉)
Swiss Centre for Affective Sciences, University of Geneva, Rue des Battoirs 7, 1205 Geneva, Switzerland
e-mail: Klaus.Scherer@unige.ch

Present Address:
S. Patel
Seton Hall University, South Orange, NJ, USA

measured and make a substantial contribution to a more comprehensive assessment of the process of emotion inferences from vocal expression.

Keywords Perception · Emotional expression · Actor portrayals · Prosody · Voice quality

Introduction

The expression and recognition of emotion through face and voice is a central domain of nonverbal communication research (Hall and Knapp 2013). In a recent, comprehensive overview of the experimental research results to date, Scherer et al. (2011; Table 2) have documented the rather high recognition accuracy for six major emotions: on average 62.5 % for dynamic facial (video) and 59.0 % for vocal (audio) expressions, largely exceeding chance levels.[1] While, on average, the recognition accuracy is similar for facial and vocal expression, there are sizeable differences between emotions—happiness and disgust are better recognized in the face, sadness and anger in the voice (see Table 2 in Scherer et al. 2011). It is interesting to note that this seems compatible with the frequent finding that it is difficult to find obvious acoustic parameters distinguishing valence differences between emotions in the voice whereas acoustic markers for arousal differences abound (e.g., tempo, amplitude and pitch variation; see Banse and Scherer 1996; Scherer 2003).

In order to understand the underlying perception and inference mechanisms, information about facial and vocal markers (cues) of specific emotions and about the use of such cues in observer's perception and inference are required. This is illustrated by the *Tripartite Emotion Expression and Perception* (TEEP) model (see Fig. 1), recently proposed by Scherer (2013) based on earlier suggestions to use a modified Brunswikian lens paradigm for the study of the nonverbal communication process (Scherer 1986, 2003). The model focuses on the communication of emotion through nonverbal cues in face, voice, body, or musical instruments, providing a framework to empirically assess cue validity and observer perception capacity.

An essential requirement to use this model for empirical research on emotion is the reliable measurement of both objectively measured distal markers or cues in face, voice, and speech of the sender and the assessment of the subjective proximal percepts in the observer, reflecting the perception and inferential use of available (or imagined) cues. In this article, we will focus on the issue of reliable assessment through a standard instrument of the proximal percepts in the vocal communication of emotion.

As software for the objective digital extraction of distal acoustic cues such as amplitude, fundamental frequency (pitch), or spectral energy distribution have become more readily available (see Juslin and Scherer 2005), there is now a body of studies demonstrating that individual emotions can be characterized by configurations of acoustic cues (see reviews by Juslin and Laukka 2003; Patel and Scherer 2013; Scherer 2003). Despite early suggestions to measure listeners' *perception* of voice and speech features that characterize the

[1] Mean values for the case of Western encoders and decoders. The recognition accuracy for expressions in static photos of facial expressions reaches 77.8 %, probably due to highly prototypical facial muscle configurations (often explicitly specified to the actors). However, these static facial stimuli cannot be reasonably compared to necessarily dynamic vocal stimuli which is why we report only the mean value for the few studies that investigated dynamic video stimuli of facial expression.

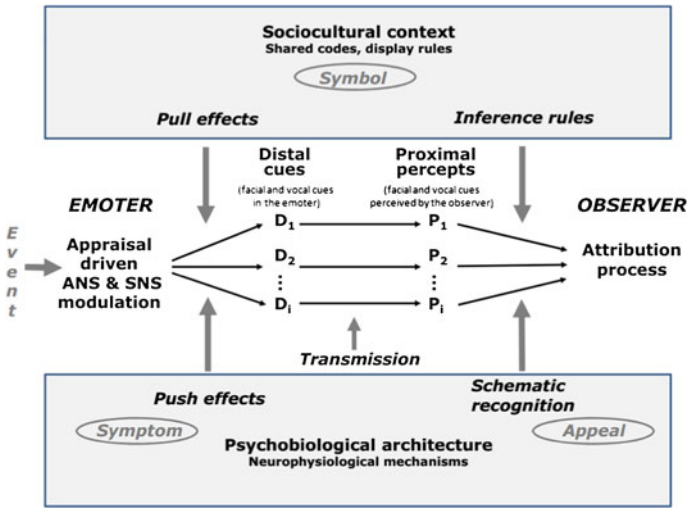

Fig. 1 The tripartite emotion expression and perception (TEEP) model, combining elements from Brunswik's lens model and Bühler's Organon model (adapted from Scherer 2013, Figure 5.5)

vocal expression of different emotions (Davitz 1964, p. 26) and an early attempt to use voice percept in a Brunswikian analysis of the recognition of personality in the voice (Scherer 1978), only very few studies have focused on ratings of voice percepts involved in the vocal communication of emotion (but see van Bezooijen (1984) who examined a large range of perceived vocal features in emotional speech using trained observers). Reasons for this neglect can be found (a) in the difficulty to reliably assess listeners' conscious perception of vocal features (e.g., Kreiman and Gerratt 1998, were unable to obtain satisfactory agreement on scales such as *roughness* or *breathiness*) and (b) in the absence of a consensual model of the mechanisms underlying the effects of emotion on voice and speech. Laver's production model (1980) was used by van Bezooijen (1984) and by Kreiman and Gerratt (1998), but the vocal dimensions were selected mostly to study individual differences in voice quality or pathological voices. A more recent proposal (Henrich et al. 2008) focuses specifically on the auditory assessment of the lyrical singing voice and performance. However, vocal features involved in vocal pathology or the singing voice cannot readily be applied to the assessment of vocal emotion expressions in normal speech. Thus, there is an urgent need to establish a standardized list of voice features that (a) can be reliably rated by non-expert raters and (b) that are likely to be crucially involved in emotional expression and communication in everyday speech. The research reported in this article was designed to address this need.

Over many years, our laboratory, in collaboration with several speech experts (see Sangsue et al. 1997)[2] addressed those issues by developing and validating a number of ratings scales, including adjectives describing voice quality—for example, "rough", "nasal", "sharp"—(a total of 13 scales) and speech characteristics—for example, "monotonous", "modulated" or "hesitant"—(a total of 6 scales). Based on the results of observer ratings of emotional and non-emotional speech using three-point Likert scales for

[2] The development of the French voice rating scales was based on earlier collaborative work for an English scale with Lou Boves and Renée van Bezooijen.

all 19 scales, the authors concluded that their preliminary results demonstrated the feasibility of the approach but that the scales required further development (Sangsue et al. 1997). Further developments and tests of the rating scales were carried out as part of an unpublished dissertation (Bänziger and Scherer 2003; Bänziger 2004) and led to the selection of scales and procedures described in this article (for Study 1).

Following this line of research, we selected eight rating scales for voice and speech percepts based on the two general criteria outlined above: (1) including rating scales that can be *reliably* used by naive listeners to assess proximal voice and speech percepts in emotional speech for a wide range of different emotions, (2) assessing the extent to which vocal *emotion expressions* can be differentiated by ratings collected with this set of scales.

In addition, we wanted to examine (1) to what extent these voice percepts can be predicted by a set of standard acoustic measures used in the vocal emotion expression literature, and (2) to what extent voice percepts and acoustic parameters, respectively, allow us to correctly classify target emotions (using multiple discriminant analyses).

We address those goals by describing the development of a set of eight scales to measure perceived voice features in emotional speech: (1) assessing the reliability of the ratings via inter-rater agreement and examining the consistency of the results across two independent studies, (2) examining the relationship between voice ratings and acoustic characteristics of the VEEs with stepwise regression analyses, and (3) assessing the emotion discrimination capacity of the voice rating scales and the acoustic parameters respectively, with the help of multiple discriminant analyses.

The vocal expression samples used for this purpose are drawn from two earlier two studies with enacted VEEs. The first study includes the data presented in the unpublished dissertation of Bänziger (Bänziger 2004; Bänziger and Scherer 2003; note that Banse and Scherer 1996, reported results for a different set of VEEs extracted from the same corpus). The second study constitutes a replication, using the same rating scales with a different set of portrayed VEEs (the Geneva Multimodal Emotion Portrayal corpus, GEMEP; Bänziger et al. 2012; Bänziger and Scherer 2010) and a simplified rating procedure. Using different groups of professional actors and different production languages allows examining the degree of generalization of the results. The results are expected to contribute to the development and validation of a new standardized instrument—the Geneva Voice Perception Scales (GVPS)—for the assessment of perceived voice and speech features. In the present article we describe the core elements to the validation of the scales with respect to reliability and usability for the description of emotional speech.

Overall, the research reported here pursues both methodological and substantive empirical aims: on the one hand (1) the development of a standard set of voice percept scales for use with models like TEEP, and on the other hand, the first attempts to understand (2) the ways in which voice percepts are anchored in, and can be predicted by, distal acoustic cues, (3) the relative power of distal and proximal cues to discriminate a set of emotions, and (4) the degree of stability of the underlying mechanisms across different experimental contexts and their comparability across languages and cultures.

Method

Selection of Stimuli (Vocal Emotional Expressions: VEEs)

A subset of VEEs were selected from two corpora of enacted emotion portrayals. The corpora and the selection criteria are described in the following.

Study 1

The VEEs used in study 1 were taken from a corpus of emotion portrayals (enacted emotional expressions) produced by professional actors in Munich, Germany, and described in several earlier publications (Banse and Scherer 1996; Scherer and Ellgring 2007a, b). For the current study, 144 expressions were chosen from this dataset. Those include 68 expressions that are also included in the data presented in earlier publications (i.e., a new selection was performed for this study also including expressions that have not been described in the earlier publications). The expressions selected were produced by nine professional actors: four men and five women. The actors were all native German speakers. The selected expressions include eight emotion categories which are reported in Table 1. For each actor and each emotion, two expressions were selected randomly among eight possible candidates, but with the constraint to include the two standard sentences produced by each actor when portraying each emotion. The two standard sentences were: (1) hät sandig prong niu ven tsie, (2) fi gött leich jean kill gos terr (pseudo-speech sentences, composed of meaningless syllables).

Study 2

The VEEs used in Study 2 were selected from a corpus of emotional portrayals produced by professional actors in Geneva, Switzerland. The Geneva Multimodal Emotion Portrayals (GEMEP) database was described in Bänziger and Scherer (2010). In the current study, 160 expressions were selected. The selection criteria matched the criteria described for Study 1 but ten actors were included (five women, all actors were French-speakers living in the Geneva area at the time of the recording, but not all native from the area). For each actor, 16 portrayals were selected, corresponding to the eight emotion categories reported in Table 1 and two standard sentences: (a) ne kal ibam soud molen(!), (b) koun se mina lod belam(?). The sentences include only a limited number of phonemes with similar realizations in most European languages. Both sentences were constructed to include the same phonemes.

Acoustic Analyses and Selection of Acoustic Parameters

All selected portrayals in Study 1 and in Study 2 were acoustically analyzed using PRAAT (open access software developed by Boersma and Weenink 2012) to extract a set of standard acoustic parameters.[3]

Study 1

Fundamental frequency (F0) was extracted using PRAAT's auto-correlation algorithm. A "conservative" manual correction of the F0 contour was performed. Detection errors were corrected where the algorithm detected periodicity in unvoiced parts of the signals. The recordings were manually segmented so as to identify pauses (speech interruptions), as

[3] The scripts used for the acoustic analyses can be downloaded at the following address: http://www.affective-sciences.org/gemep/perceived_voice. Further supplementary materials (audio examples and details of statistical results) are also available at the same address.

Table 1 Emotion categories selected in both data sets

Arousal level	Emotion family	Original label for the actors/senders		Translation
		Study 1	Study 2	(short label in parenthesis)
Low	Anger	Kalter Ärger	Irritation	Cold anger, irritation (irrit.)
High	Anger	Heisser Ärger	Colère	Hot anger (anger)
Low	Happiness/joy (positive emotion)	Stille Freude	Plaisir	Quiet joy in Study 1 (happi.)/ pleasure in Study 2 (pleas.)
High	Happiness/joy (positive emotion)	Überschäumende Freude	Joie	Excited joy, elation (elat.)
Low	Sadness	Stille Trauer	Tristesse	Sadness (sad.)
High	Sadness	Verzweiflung	Désespoir	Despair (desp.)
Low	Fear	Angst	Inquietude	Anxiety (anx.)
High	Fear	Panische Furcht	Peur panique	Panic fear (panic)

All emotions in Study 2 were defined to correspond to the categories used in Study 1. The difference in labels is due to the translation from German to French. There is one exception for quiet joy and pleasure which did not correspond to the same definition but were the closest "match" in both datasets

well as voiced and unvoiced segments. Several parameters were extracted from the F0 and several absolute and relative duration parameters were computed for different speech segments and pauses. Further parameters were extracted from the intensity contour. The proportion of spectral energy in various frequency regions of the long term averaged spectrum (LTAS) was also investigated. The spectrum was segmented into bands, matching the approach and results reported by Banse and Scherer (1996). Spectral parameters were extracted separately for the voiced and the unvoiced parts of the expressions. In total, 44 parameters were extracted from the signals, many of which were very highly intercorrelated. All parameters were independently standardized within speaker in order to control for variations due to inter-individual differences (using z-transformations).

In order to reduce multicollinearity in subsequent analyses, we selected a smaller set of parameters on the basis of an exploratory principal component analysis of the 44 extracted parameters. The principal component analysis yielded a data structure with nine components accounting for 80 % of the variance in the data. One optimally representative parameter was selected for each component and the mean acoustic intensity which loaded on several components was added to this set leading to a total of ten selected acoustic parameters which are listed in Table 2.

Study 2

The parameters described in Study 1 were extracted for the VEEs used in Study 2. However, no manual corrections of F0 or duration were performed in Study 2, as the effects of the manual corrections in Study 1 were estimated to be negligible. Several measures on the long term average spectrum (LTAS) included in Study 1 (various spectral bands on the voiced and the unvoiced spectrum) were not used in the analyses, since they did not make significant independent contributions to the differentiation of emotions in Study 1. In addition it appeared that some additional parameters could be reliably extracted. Shimmer, jitter, and harmonics-to noise ratio (HNR) were extracted and added to

Table 2 Selected acoustic descriptors in Study 1 and in Study 2

Domain	Description	Label	Used in Study 1	Used in Study 2
Fundamental frequency (F0)	Minimum	F0.min	X	
	5th-percentile	F0.p05		X
	Range (difference between minimum and maximum)	F0.range	X	X
Intensity	Mean	Int.mean	X	X
	Range (difference between minimum and maximum)	Int.range	X	X
Duration	Total duration (of the utterance)	Dur.tot	X	X
	Relative duration of voiced segments on speech segments (duration of voiced divided by the sum of the duration of voiced and unvoiced segments, i.e. excluding phonetic interruptions)	Dur.v/art	X	X
Distribution of energy in the LTAS, voiced segments only	0–1,000 Hz (relative to 0–8,000 Hz)	LTSv < 1,000	X	
	300–500 Hz (relative to 0–8,000 Hz)	LTSv.500	X	
	600–800 Hz (relative to 0–8,000 Hz)	LTSv.800	X	
Distribution of energy in the LTAS, unvoiced segments only	0–1,000 Hz (relative to 0–8,000 Hz)	LTSn < 1,000	X	
Irregularity of voicing	Harmonics-to-noise-ratio	HNR		X

Different parameters are used in Study 1 and in Study 2 because of initial differences in choices of parameter extraction (some parameters extracted in Study 1 were not extracted in Study 2 and some additional parameters were added in Study 2). The selection of a limited number of parameters for further analyses was derived empirically, based on the amount of variance shared among the extracted parameters. This was done independently for Study 1 and for Study 2

the parameter set. All parameters were standardized separately within speakers (using z-transformations).

Given that a modified list of parameters and a modified procedure for the extraction were used, we again performed a principal component analysis of the complete parameter set in Study 2 to select a smaller set of parameters with reduced collinearity. The principal component analysis performed on 34 extracted parameters produced six components accounting for 86 % of the variance in the data. Again, one representative parameter was selected for each component. In addition, a measure of F0 floor (the value of the 5th percentile of the F0 values extracted for each portrayal) was added to the parameter set.[4] The list of the seven selected parameters is presented in Table 2.

[4] This parameter (F0 floor, see Tolkmitt and Scherer 1986) was added because of the assumption that although pitch and intensity are highly correlated in both Study 1 and Study 2, they may still constitute relevant and partly independent aspects of vocal communication of emotion. We included a measure of average acoustic intensity in Study 1 using the same rationale.

Ratings: Selection of Rating Scales

Study 1

The selection of rating scales was based on the work reported in Sangsue et al. (1997), identifying potentially relevant terms (from a list of French adjectives), which might be relevant for the characterization of emotional voice and speech. In a series of tests reported in an unpublished dissertation, Bänziger (2004) further investigated the aspects that could be rated by untrained listeners using French nouns and adjectives to designate vocal dimensions. Those tests essentially showed that naive raters (students in psychology with no specific education in speech/voice analysis or description) agreed on only few terms that could be used to describe speech or voice quality without referring to "external" factors such as age, emotions, or personality. This observation suggested decreasing the number of scales to nine scales, and later to eight scales, for which naive raters appeared to share a common understanding of the designated voice/speech characteristics. Tests showing that most people disagreed even on the definition of simple voice qualifiers (such as roughness or sharpness) further lead to asking an experienced speaker (a collaborator on a research project on prosody and amateur singer) to produce extreme examples to illustrate the contrasts involved by high and low levels of the rating scales under consideration. The final selection of scales used to collect ratings in both studies presented in this article is described in Table 3.

Study 2

The scales used in Study 1 were used again in Study 2 for ratings of different VEEs and with a new group of raters, using a simplified rating procedure.

Ratings: Participants (Raters)

Study 1

As the rating procedure used in Study 1 (see below) was more time-consuming than a conventional rating procedure, four groups of listeners were recruited to evaluate subsets of the 144 VEEs included in Study 1. The groups were composed of 15–16 first-year students in psychology at the University of Geneva. All raters had normal hearing capacity and participated in the study against course credit. Students were randomly allocated to one of four groups. Group 1 consisted of 14 women and two men (average age = 21.3 years, $SD = 4.3$); group 2 consisted of 10 women and five men (average age = 21.7, $SD = 4.3$); group 3 consisted of 13 women and two men (average age = 20.2, $SD = 1.7$); group 4 consisted of 11 women and four men (average age = 21.4, $SD = 6.1$). The study took place in a small laboratory for psychological assessment at the University of Geneva.

Study 2

Based on the results of Study 1, a simplified rating procedure (see below) was used in Study 2, involving nineteen raters (10 women and 9 men with an average age of 22.4 years, $SD = 2.2$) who were asked to assess all 160 portrayals in one rating session. The raters received a financial compensation of 60 CHF for their contribution. All participants had normal hearing. The study took place in a small laboratory for psychological assessment at the University of Geneva.

Table 3 Rating scales used in both studies

English translation		French scale names (used in the study)	
Scale	Direction	Scale	Direction
Pitch	Low ↔ high	Hauteur	Grave ↔ aiguë
Loudness	Weak ↔ strong	Volume	Faible ↔ forte
Intonation	Monotonous ↔ accentuated	mélodie	Monotone ↔ modulée
Speech rate	Slow ↔ fast	Vitesse	Lente ↔ rapide
Articulation	Poor ↔ good articulation	Articulation	Mal ↔ bien articulée
Instability	Steady ↔ trembling	Stabilité	Ferme ↔ tremblante
Roughness	Not rough ↔ rough	Qualité rauque	Non rauqe ↔ rauque
Sharpness	Not sharp ↔ sharp	Qualité perçante	Non perçante ↔ perçante

Pitch, loudness, instability, roughness, and sharpness referred explicitly to voice in the ratings studies, while intonation, speech rate, and articulation referred explicitly to speech. The ninth scale was "fluidity" of speech (defined as a dimension ranging from absence of hesitations or interruption in the speech-flow to many hesitations or interruptions). Our data showed that this dimension could be reliably rated, but was considered as not relevant for the very short VEEs used in our studies (pseudo-speech sentences consisting of only 6–7 syllables)

Ratings: Procedure

The ratings for Study 1 and for Study 2 were collected several years apart using different procedures to collect the ratings.

Study 1

Kreiman and Gerratt (1998) showed that the evaluation of vocal quality on scales such as "rough" or "breathy" do not yield reliable judgments (inter-rater reliability and test–retest reliability were found to be low). According to these authors, internal standards of comparison (anchors) used by listeners when they are making judgments vary from one listener to another and also vary over time for a single listener. In order to address the problem of variable anchors, we adapted a rating procedure introduced by Granqvist (1996), in which all expressions produced by one speaker can be rated simultaneously and direct comparison is used to ensure that the standard for comparison is not fluctuating for a given speaker. Using this approach, a visual analog scale was presented to the listeners on a computer screen. The task of the listeners was to place the VEEs on this scale. All expressions produced by a given speaker appeared on the screen in random order, as identical icons, which could be played by double-clicking on the icons. The raters' task was to place them on the scale depending on the value he or she allocated to each recording. Listeners were free to listen to the VEEs again as often as they wished, and could modify their answers. The VEEs produced by different speakers were presented on successive screens, so that the judgments were relative to the range of variation of a given speaker (and insensitive to inter-speaker differences). In addition, two recordings illustrating the ends of each scale were presented at the bottom of the screen.[5] Those recordings were

[5] Those recordings are all produced by one speaker (a research collaborator and expert in speech analysis). They are not used as anchors for the ratings, but only to illustrate the specific contrast represented in every scale (i.e., they help to define the scales for the raters). It was clear to the raters that—for example—the pitch level of the emotional expressions was not to be compared directly with the pitch range given by the examples. Those recordings can be accessed along supplementary materials on the website indicated in footnote 3.

presented as illustrations and were not meant to be used as anchors. Pre-tests of the procedure showed that listeners understood the procedure and were able to use it without difficulty.

Each rater evaluated 48 recordings (2 expressions × 8 emotions × 3 speakers) on the eight scales described in Table 3 (for a total of 48 × 8 = 384 ratings). Identical computers, sound cards and headphones were used for all participants and all sessions. The scales and the speakers were presented sequentially, in a different random order for each listener. Answers were recorded by the computer on a continuous scale from 0 to 10 (no numbers were visible to the raters; labels described in Table 3 were indicated as scale endpoints).

Four groups of raters were recruited to assess the total set of VEEs included in this study (144 expressions produced by 9 actors). The raters in the four groups assessed the VEEs produced by one common speaker in order to control for group differences. No systematic differences were found across groups. In order to keep the number of ratings (and hence the reliability of the assessments) comparable across speakers we randomly removed 75 % of the ratings collected for the speaker who's VEEs had been evaluated by all groups. We then computed average ratings for each VEE ($N = 144$) based on either 15 or 16 ratings for each scale.

Study 2

In Study 2, a traditional rating procedure with visual analog scales was used to examine the possibility of obtaining reliable ratings at a lesser cost. The ratings were collected in eight successive blocks for separate scales (the list of scales is provided in Table 3 and was the same as in Study 1). Breaks were allowed in between rating blocks. The order of the blocks (i.e., scales) was randomized for each participant. The scales were shown in visual analog format (i.e., without numeric values, only labels defining the scale end points). The audio illustrations of each scale were made available on both ends of the scale (same illustrations as in Study 1, see footnote 5). Participants were required to listen to the examples before rating the portrayals on each new scale. In each block, all samples were randomly presented by speaker. The speaker order was randomized across blocks, and the order of the blocks was randomized. A replay button was added to the bottom of the screen that allowed the VEEs to be replayed. Participants listened to each portrayal and then reported their answer immediately after hearing each expression on the visual-analog scale. Answers were recorded as values ranging from 0 to 100 (no numbers were shown to the raters). Identical computers, sound cards and headphones were used for all participants.

Participants provided ratings in one session of 3 h.[6] Each participant provided 160 × 8 = 1,280 ratings for this part of the rating study.

Results

The results reported for both studies are organized according to the goals listed in the introduction: (1) assessment of inter-rater agreement, providing an estimate of the

[6] Additional ratings were collected (concerning the emotions perceived in the recordings) from the same raters in Study 2. Those additional ratings are not described in the current paper and were collected in a second step; i.e., a new set of instructions was presented to the raters after the procedure described in the current article and the recordings were replayed in a new random order for the collection of additional ratings.

reliability of the ratings; (2) regressions of voice ratings on acoustic parameters to assess the extent to which the voice ratings can be predicted by a set of acoustic parameters; and (3) using discriminant analysis to determine the relative accuracy with which the expressed emotions can be classified on the basis of the distal acoustic cues and the proximal voice percept ratings, respectively. In order to facilitate the examination of the convergence of the results in the two studies, we will report both sets of results jointly under each of these headings.

Reliability (Inter-Rater Agreement) and Collinearity

Study 1

Inter-rater reliability was estimated for the eight scales and for each group of raters in the form of intra-class correlations (ICC). The "single measure" ICC-r is an estimate of the average correlation between ratings of all raters. The "average measure" ICC-R is equivalent to Cronbach's alpha estimate (Rosenthal 1987). There were no significant differences across groups. To simplify the data presentation we report average values for the four groups in Table 4 (columns 1 and 2).

Study 2

Inter-rater reliabilities (ICC-*r* and ICC-*R*) were computed for the eight scales and are reported in Table 4 (columns 3 and 4). The reliability estimates were comparable in both studies despite several methodological differences (different selection procedures for the portrayals, different procedures used to collect the ratings).

Inter-rater agreement obtained with the more complex rating procedure used in Study 1 was not noticeably different from habitual levels, suggesting that a classic rating procedure may be sufficient for the purpose at hand. We also note that there are sizable differences in reliability estimates across scales, with roughness on the lower end (reliability estimate .86 in Study 1 and .84 in Study 2; average inter-rater correlation estimate .31 in Study 1 and

Table 4 Reliability of the ratings in Study 1 and Study 2, intraclass correlations (ICC), r = single measure, R = average measure, for eight rating scales

Voice scales	Study 1		Study 2	
	ICC-*r*	ICC-*R*	ICC-*r*	ICC-*R*
Roughness	.305	.858	.216	.840
Articulation	.316	.865	.339	.907
Intonation	.395	.907	.375	.919
Instability	.468	.930	.427	.934
Pitch	.517	.942	.554	.959
Sharpness	.588	.952	.579	.963
Speech rate	.661	.967	.605	.967
Loudness	.854	.989	.811	.988

There were on average 15 raters in Study 1 and 19 raters in Study 2. The ICC-*R* estimates are therefore slightly better only because of a larger number of raters. ICC-*r* are more comparable estimates with a different number of raters

.22 in Study 2) and loudness on the upper end (reliability estimate .99 in both studies; average inter-rater correlation estimate .85 in Study 1 and .81 in Study 2). We conclude that the set of scales allows reliable measurement, at habitual levels of ICCs, with a classic rating format using visual analog scales (as used in Study 2).

In both studies, an average rating score was computed for each emotion portrayal on each scale. In order to assess the degree of dependence between the rating scales, Pearson correlation coefficients and principal components analyses (PCA) were computed. The respective tables and discussions can be found in supplemental materials, Section A (the document with supplemental results can be downloaded on the website indicated in footnote 3).

In both studies, the correlations are largest among the scales intonation, pitch, loudness and sharpness, constituting a strong first factor in the PCAs. This pattern of correlations probably reflects vocal effort, an emotion-relevant voice production factor frequently reported and discussed in studies of VEEs (Scherer 2003; Sundberg et al. 2011). It should be noted that, as our selection of VEEs includes large variations in emotional arousal (half of the expressions were chosen to represent low-arousal and the other half high-arousal in both studies), correlations between cues that reflect vocal effort will be automatically boosted due to the nature of the distribution on the bipolar continuum. As the collinearity patterns are highly dependent on study design, especially the choice of the number and types of emotions, we will not discuss these results in further detail as they may not generalize to studies involving less extreme variation in emotional arousal.

Relationship Between Voice Ratings and Acoustic Measures: Regression Analyses

The second goal formulated in the introduction concerns the relationship between the voice percept ratings and acoustic measures extracted from the VEE samples; in particular the extent to which the values collected with the voice rating scales can be accounted for by acoustic parameters that are more regularly used for the description of emotional speech. Using voice ratings is more relevant if the values obtained are not totally redundant with acoustic descriptors. It is especially interesting to observe if different voice ratings scales can be better accounted for by pertinent acoustic measures (e.g., perceived loudness by acoustic indicators of vocal effort).

The correspondence between a selection of acoustic descriptors (10 acoustic parameters selected in Study 1 and 7 acoustic parameters selected in Study 2, see the methods section on the selection of acoustic parameters) and the voice ratings were estimated using stepwise regressions.

Study 1

Ten acoustic parameters (list in Table 2) were entered as predictors of perceived vocal dimensions in multiple regressions. Stepwise regressions were used to select the best predictors for each rating scale. The results are reported in Table 5. The eight regression models for Study 1 are shown on the left side of this table. For each regression, the effect size (R^2) and significance test are shown first, followed by the individual contributions of the variables (acoustic parameters) entered in each model. The selected acoustic parameters explain a sizeable proportion of the variance for the following ratings scales: loudness (with 88 % of variance explained), sharpness (87 %), speech rate (79 %), intonation with (67 %), and pitch (65 %). They account less well for the scales instability (35 %), quality of articulation (32 %), and roughness (28 %).

Table 5 Stepwise regressions of voice ratings on selected acoustic parameters in Study 1 and in Study 2

Vocal dimension	Study 1				Study 2			
Loudness	$R^2 = .882$ $F(3,140) = 348.417$, $p < .001$				$R^2 = .954$ $F(2,157) = 1{,}621.934$, $p < .001$			
	Acoustic parameters	β	*t*	*p*	Acoustic parameters	β	*t*	*p*
	Int.mean	.724	15.973	.000	Int.mean	.947	44.157	.000
	Int.range	.178	5.496	.000	Int.range	.049	2.262	.025
	LTSv < 1,000	−.160	−3.764	.000				
Sharpnesss	$R^2 = .867$ $F(5,138) = 180.511$, $p < .001$				$R^2 = .935$ $F(3,156) = 749.349$, $p < .001$			
	Acoustic parameters	β	*t*	*p*	Acoustic parameters	β	*t*	*p*
	Int.mean	.510	9.138	.000	Int.mean	.761	21.422	.000
	F0.range	.253	6.282	.000	F0.range	.118	4.996	.000
	F0.min	.185	5.071	.000	F0.p05	.180	5.521	.000
	LTSv < 1,000	−.150	−3.280	.001				
	Int.range	.075	2.114	.036				
Speech rate	$R^2 = .786$ $F(4,139) = 127.579$, $p < .001$				$R^2 = .553$ $F(4,155) = 47.870$, $p < .001$			
	Acoustic parameters	β	*t*	*p*	Acoustic parameters	β	*t*	*p*
	Dur.tot	−.573	−13.322	.000	Dur.tot	−.699	−11.952	.000
	Int.mean	.434	6.280	.000	Int.range	.276	4.019	.000
	v.0–1 k	−.181	−3.111	.002	F0.p05	.251	3.998	.000
	Dur.v/art	−.148	−2.974	.003	F0.range	.119	2.040	.043

Table 5 continued

Intonation	$R^2 = .669$ $F(6,137) = 46.088$, $p < .001$				$R^2 = .676$ $F(4,155) = 81.029$, $p < .001$			
	Acoustic parameters	β	*t*	*p*	Acoustic parameters	β	*t*	*p*
	F0.range	.396	5.853	.000	F0.range	.363	6.746	.000
	Int.mean	.213	2.507	.013	Int.mean	.342	4.281	.000
	Int.range	.209	3.534	.001				
	F0.min	.190	3.246	.001	F0.p05	.231	3.138	.002
	LTSn < 1,000	−.147	−2.842	.005	Hnr	.218	4.525	.000
	Dur.tot	−.130	−2.230	.027				
Pitch	$R^2 = .647$ $F(3,140) = 85.534$, $p < .001$				$R^2 = .669$ $F(3,156) = 105.287$, $p < .001$			
	Acoustic parameters	β	*t*	*p*	Acoustic parameters	β	*t*	*p*
	F0.min	.408	7.110	.000	F0.p05	.624	12.424	.000
	F0.range	.402	6.248	.000	F0.range	.311	6.358	.000
	Int.mean	.250	3.492	.000	Hnr	.174	3.593	.000
Instability	$R^2 = .351$ $F(3,140) = 25.215$, $p < .001$				$R^2 = .235$ $F(3,156) = 15.979$, $p < .001$			
	Acoustic parameters	β	*t*	*p*	Acoustic parameters	β	*t*	*p*
	F0.min	.485	6.564	.000	F0.p05	.731	6.544	.000
	Dur.tot	.346	4.880	.000	Int.mean	−.613	−5.025	.000
	LTSv < 1,000	.235	3.071	.003	F0.range	.243	3.012	.003

Table 5 continued

Articulation	$R^2 = .324$ $F(5,138) = 13.223, p < .001$				$R^2 = .343$ $F(5,154) = 16.056, p < .001$			
	Acoustic parameters	β	*t*	*p*	Acoustic parameters	β	*t*	*p*
	Int.mean	.468	4.555	.000	Int.mean	.922	7.947	.000
	F0.min	−.392	−4.782	.000	F0.p05	−.694	−6.568	.000
	F0.range	−.303	−3.294	.001	F0.range	−.194	−2.451	.015
	Int.range	.310	3.760	.000	Dur.tot	.224	3.250	.001
	LTSn < 1,000	−.200	−2.707	.008	Hnr	.140	1.978	.050
Roughness	$R^2 = .278$ $F(4,139) = 13.407, p < .001$				$R^2 = .330$ $F(3,156) = 25.611, p < .001$			
	Acoustic parameters	β	*t*	*p*	Acoustic parameters	β	*t*	*p*
	LTSv < 1,000	−.346	−3.267	.001	Hnr	−.417	−6.062	.000
	LTSn < 1,000	.258	3.492	.001	F0.p05	−.296	−4.367	.000
	Dur.tot	.221	2.829	.005	Dur.tot	.240	3.595	.000
	Int.mean	.231	2.089	.038				

β = standardized beta weight

Study 2

Seven acoustic parameters (list in Table 2) were entered as predictors of perceived vocal dimensions in stepwise regressions. The results are reported in Table 5 (on the right side of this table). The results in Study 1 were largely replicated, despite the many methodological differences listed earlier (different VEEs, different raters, different ratings procedures) and despite using partly different acoustic parameters in both studies. Again, more of the variance can be accounted for in the case of the scales: loudness, sharpness, speech rate, intonation and pitch than for the scales instability, quality of articulation, and roughness. However, we observe one difference: for the overall variance explained for perceived speech rate for which the acoustic parameters entering the regression equation in Study 1 explained more of the variance than the acoustic parameters selected in Study 2.

In summary, acoustic measures account for very different amounts of the variance in voice ratings, depending on the specific rating scale that is considered. These discrepancies appear to be consistent even when using different emotion portrayals and partly different acoustic parameters.

Emotion Discrimination Via Voice Ratings and Acoustic Parameters

The third goal of this paper was to demonstrate the distinctiveness of the voice ratings and the acoustic cues for different expressed emotions by examining the extent to which the voice ratings and, comparably, the acoustic cues could discriminate the eight emotion categories included in both studies using multiple discriminant analyses (MDAs). The average voice ratings used had yielded significant effects for an Emotion factor in repeated measures ANOVAs (the details of these ANOVAs are reported in Section B1 of the supplemental materials, to be downloaded at the address reported in footnote 3). Because the classification would be improved by the inclusion of a larger number of relevant predictors, we decided to use only eight out of the ten acoustic parameters selected in Study 1 (i.e., a number equivalent to the number of voice ratings used). We removed the two acoustic parameters (LTSv.500 and LTSn < 1,000) that did not show significant effects for the Emotion factor in repeated measures ANOVAS.

Study 1

A first MDA was computed using the values obtained for the eight selected acoustic parameters. The categories used in the analysis are the eight expressed emotions (see Table 1). Chi square tests on Wilks' lambda indicate that the first three discriminant functions make statistically significant contributions to the discrimination ($p < .05$). The parameters int.mean (.849) and LTSv < 1,000 (−.542) load on the first function. F0.min (.726) loads on the second function. Dur.tot (.687) loads on the third function. LTSv.800 (.607), F0.range (−.550), int.range (.824) and dur.v/art (−.576) load respectively on functions 4–7 (which do not make significant contributions to the discrimination). The discriminant functions achieve a level of 60 % correct classification (50 % with "leave-one-out" cross-validation); see Table 6, column 2 for the percentages of correct classification by emotion.

A second MDA was computed using the averaged values obtained for the 144 emotion portrayals on the eight voice rating scales in Study 1. The categories used in the analysis are again the eight expressed emotions (see Table 1). Chi square tests on Wilks' lambda indicate that the first five discriminant functions make statistically significant contributions

Table 6 Classification accuracy based on discriminant functions extracted with acoustic parameters or with voice ratings (in Study 1 and in Study 2)

	Study 1		Study 2	
	Acoustic	Ratings	Acoustic	Ratings
Irritation	61.1	66.7	60.0	65.0
Anxiety	61.1	50.0	50.0	55.0
Sadness	72.2	88.9	90.0	85.0
Happiness/pleasure	66.7	61.1	75.0	70.0
Hot anger	83.3	77.8	90.0	100.0
Panic fear	66.7	77.8	80.0	95.0
Despair	50.0	61.1	55.0	70.0
Elation	22.2	72.2	55.0	80.0
Mean	60.4	69.5	69.4	77.5

to the discrimination ($p < .05$). The scales loudness (.791), sharpness (.781) and intonation (.694) load on the first function (loadings/correlations are indicated in parenthesis). Instability (.867) and pitch (.509) load on the second function. Roughness (−.666) loads on the third function. Speech rate (.600) loads on the fourth function and articulation (.696) on the fifth function. The discriminant functions achieve a 69 % correct classification (61 % with "leave-one-out" cross-validation). The percentages of correct classification by emotion are shown in Table 6, column 3. The complete confusion matrices are reported in supplementary materials (section B2, to be downloaded from the website indicated in footnote 3).

Study 2

Again, an MDA was computed using the values obtained for the seven selected acoustic parameters in Study 2 (see Table 2). The categories used in the analysis are the eight expressed emotions (see Table 1). Chi square tests on Wilks' lambda indicate that the first four discriminant functions make statistically significant contributions to the discrimination ($p < .05$). The parameter int.mean (.951) loads on the first function. Dur.tot (.668) loads on the second function. F0.p05 (.686) loads on the third function. Int.range (−.744) loads on the fourth function. F0.range (.699) loads on the fifth function. HNR (.701) and dur.v/art (.650) load on the sixth function. The discriminant functions achieve a 69 % correct classification (61 % with "leave-one-out" cross-validation). The percentages of correct classification by emotion are shown in Table 6, column 4.

A final MDA was computed using the averaged values obtained on the eight voice rating scales for the 160 VEEs included in Study 2. The categories used in the analysis are the eight expressed emotions. Chi square tests on Wilks' lambda indicate that the first four discriminant functions make statistically significant contributions to the discrimination ($p < .001$). The fifth function fails to pass the customary 5 % threshold for type A error ($p = .081$) The scales loudness (.919) and sharpness (.873) load on the first function. Instability (.688) loads on the second function. Speech rate (.742) loads on the third function. Articulation (.561) loads on the fourth function. Roughness (.699) and pitch (−.588) load on the fifth function. Intonation (.458) loads on the sixth function. The discriminant functions achieve a 77 % correct classification (68 % with "leave-one-out"

cross-validation). The percentages of correct classification by emotion are shown in Table 6, column 5. The confusion matrices are reported in supplementary materials (section B2, the document can be downloaded at the address reported in footnote 3).

The overall hit rate for the classification based on acoustic parameters (60 % in Study 1 and 69 % in Study 2) was much larger than a random classification (12.5 %). The two sets of acoustic parameters did not achieve consistently better classifications than the voice rating scales. The voice ratings achieve overall slightly larger classification accuracy (69 % in Study 1 and 77 % in Study 2) than the acoustic parameters. For the category elation, the classification success of the voice ratings was clearly superior to the classification success of the acoustic parameters.

The overall classification hit rate was somewhat higher in Study 2, particularly for emotions with high-arousal and also for pleasure which was slightly better classified than happiness in Study 1 (see results in Table 6). High-arousal emotions were occasionally confused with low-arousal emotions in Study 1, while this never occurred in Study 2. Discriminant functions making significant contributions were defined principally by the scales loudness and sharpness (used to define the first discriminant function in both analyses), instability (function 2 in both studies), speech rate (function 4 in Study 1 and function 3 in Study 2) and articulation (function 5 in Study 1 and function 4 in Study 2). The average emotion scores on the functions obtained with the same scales in both studies were highly correlated; $r = .92$ for average emotion scores obtained with function 1 (mainly defined by loudness and sharpness) and $r = .92$ also for scores obtained with function 2 (mainly defined by instability); $r = .87$ for emotion scores obtained with functions on which speech rate had highest loadings; $r = .78$ for emotion scores obtained with functions mainly defined by articulation. The principal difference was that roughness made an independent contribution to the discrimination in Study 1; while it did not in Study 2 (nevertheless the classification was slightly better in Study 2). The classification patterns and the underlying discriminant functions were not exactly identical across the two studies, but they can be considered as highly similar in nature and direction.

Discussion and Conclusion

The two studies presented in this article describe our efforts to develop and validate a reliable instrument for the assessment of perceived voice and speech features in with the aim of contributing to a more comprehensive account of the emotion communication process (as exemplified by the TEEP model shown in Fig. 1). In the interest of examining the generalizability of the scale, the two studies include different sets of emotion portrayals (produced by different speakers, from different language groups), which were assessed by different groups of listeners, using different rating procedures. Study 2 was not designed to be an identical replication of Study 1 and we did therefore not intend to directly compare or integrate the results of both studies. As Study 2 introduces several variations both in the material investigated and in the procedures used to collect the ratings, we were able to examine if the general findings would be affected by such variations. In this respect, we observed that despite the methodological differences, many results were consistent between both studies. We found similar reliability estimates for various voice scales, similar contrasts in scores for different emotions, and similar discrimination performance.

Regarding our first goal to assess the reliability of the ratings across listeners, we found that the ratings showed consistent inter-rater reliabilities in both studies, with large variation across scales (high agreement for some scales and somewhat lower agreement for

others). These patterns of reliabilities confirm similar results reported by van Bezooijen (1986) for perceptual voice judgments in the Dutch language (based on Laver 1980). Eight of the voice scales in the respective instrument reflect dimensions of voice description similar to the scales described in this article. More recently, Biemans (2000, Table 6.2) confirmed the reliability results obtained by van Bezooijen (1986) using lay raters. This consistency in obtaining high reliabilities for perceptual voice rating scales demonstrates the possibility to gather reliable ratings of voice and speech characteristics for most of the voice and speech characteristics considered in the studies presented above. Given the convergence of results in Study 1 and Study 2 and the similarities with the results obtained by van Bezooijen and Biemans, we expect that further studies will yield similar reliability indices for ratings of the vocal features examined here.

The second goal addressed in the two studies concerned the relationships between voice/speech ratings and acoustic parameters. A limited number of acoustic parameters were used in both studies in order to reduce the degree of collinearity between the acoustic parameters. The selected acoustic parameters accounted for a large proportion of the shared variance in the voice ratings, which probably reflects vocal effort and emotional arousal. *Pitch*, *sharpness*, *loudness*, and *intonation* were highly inter-correlated and were also largely predicted by acoustic measures of intensity, fundamental frequency and energy distribution in the long term averaged spectrum. Ratings of *speech rate* were well accounted for by duration measures in Study 1 but less well in Study 2. However, acoustic parameters did not account as well for ratings that are conceptually related to "speech/voice quality" (*instability, articulation, roughness*). This suggests that such ratings might provide descriptions of VEEs that are not captured by simple acoustic measures aggregated over sentences, pointing to the need of developing measurements for additional parameters in future research (especially given that the standard set of acoustic parameters routinely used in vocal emotion expression research was developed by phoneticians studying speech processes rather than nonverbal vocal expression).

The studies reported here, as well as most of the pertinent work in the literature, have used *aggregated* acoustic variables, reflecting intensity, fundamental frequency, and energy distribution in the long term spectrum, as well as speech rhythm over whole sentences. As suggested in earlier reviews of findings in this field, these classic acoustic parameters reflect emotional arousal but are of little importance in the communication of the valence or pleasantness dimension of emotion (Banse and Scherer 1996; Juslin and Scherer 2005; Scherer 2003). Aggregated acoustic measures most probably fail to capture important dynamic voice cues that may be used and integrated by listeners when they derive emotional attributions from speech. The results presented here clearly point to the need to further develop acoustic measures to describe the dynamic variations in speech, including measures reflecting voice and speech quality at the segmental level. Listeners' conscious ratings and descriptions of vocal expressions may help and guide such developments (for example, via detailed measurements of the dynamic acoustic properties of voice samples that receive extreme ratings on relevant scales).

The third and last goal formulated in the introduction was to investigate the effect of emotions on voice ratings and the possibility to differentiate between emotions using ratings of voice and speech characteristics. Our results showed that the voice ratings allowed discriminating the eight emotions included in both studies with a high level of accuracy, comparable or even larger than the accuracy achieved with a similar number of acoustic measures.

The results of the MDAs showed that while both sets of predictors (acoustic measures and voice ratings) produce hit rates much larger than what would be expected by chance,

the hit rates are somewhat higher when voice ratings are used. This confirms results reported by van Bezooijen (1984) who showed that direct prediction of portrayed emotions from acoustic measurements was less successful than from perceptual ratings. An interesting aspect is that the difference (in classification accuracy based on acoustic measures versus voice ratings) is most pronounced in the case of elation, an emotion that combines positive valence with high arousal and is often confused in judgment studies with hot anger or fear because of the latter. This implies that perceptual voice ratings may reflect cues that are pertinent for valence detection and that have not yet been identified in acoustic analyses. Our results suggest that this may be a promising line of investigation for further research. Scherer (1986) has hypothesized that pleasantness might be mostly expressed in changes in vocal tract shape (e.g., faucal and pharyngeal expansion or constriction, relaxation or tensing of tract walls, vocal tract shortening and lengthening). These vocal tract changes will affect the energy distribution in the spectrum and most importantly, the formant frequencies and bandwidth. In speech material with constantly changing phonemes, these effects will be buried in the long term spectrum. Analyses are required that examine potential formant effects separately for individual phonemes (see for example Tartter and Braun 1994 or Robson 1999). While such analyses will obviously require greater investment than the standard parameters used in the past, such measures might help to explain the effects found with voice ratings and thus provide important clues to identify the underlying mechanisms of valence expression and communication in the voice.

The fourth goal of this research as announced in the introduction concerned a preliminary examination of the degree of stability of the underlying mechanisms across different experimental contexts and their comparability across languages and cultures. Obviously, the degree of systematic variation of experimental context between our two studies is highly limited. Similarly, we compared speakers (actors) from two relatively similar languages and cultures (German and French). In consequence, any conclusion on this issue would be quite inappropriate and further research using more differentiated experimental designs and studying more remote, in particular non-Western, languages and cultures is clearly needed. Yet, the high degree of similarity between the two studies reported is compatible with the general tendency reported in recent reviews of the literature on emotional expression, findings increasingly pointing to a relatively high degree of universality, tempered by a discernible amount of "dialectal" differences between languages and culture (see Scherer et al. 2011).

In conclusion, the two studies presented in this paper confirm the feasibility of our approach to the assessment of perceived vocal cues in emotional communication. Our studies demonstrate that the ratings collected with the scales described in this paper are sufficiently reliable and are relevant to the description of emotional speech. We propose to refer to the set of scales used in the two studies as the "Geneva Voice Perception Scales (GVPS)". A complete description of the scales and the procedure for administration can be downloaded by interested researchers from the URL indicated in footnote 3. We believe that this new instrument can be used to gain further insight into the vocal aspects involved in the communication of emotion. In the long run, we expect that listeners' ratings can help to identify the vocal features that are relevant for emotional communication in the dynamic variation of voice quality.

We expect that this approach can be further developed and lead to a more integrated view on the vocal communication of emotion, including the study of production and perception of emotional expressions. In particular, adding vocal percepts to the study of the emotion communication process might help to identify auditory cues for certain emotions that listeners perceive but for which we have as yet no appropriate measurement

procedures on the level of acoustic parameters. This deserves high priority for future research, particularly with respect to the identification of vocal markers of valence. The results reported above clearly show that listeners are rather well able to distinguish different degrees of valence on the basis of auditory cues in isolation. In consequence, it seems fruitful to use results on vocal emotion perception to explore the nature of the respective vocal production and the respective acoustic manifestations (see Sundberg et al. 2011). Work in this direction might also provide some glimpses concerning the evolution of vocal emotion expression (see Scherer 2013) and provide a more solid theoretical background to examine differences between languages and cultures. A final desideratum, after having provided tools for the measurement for both the distal and proximal aspects of vocal emotion communication, is the development of appropriate statistical modeling tools (such as the Brunswikian lens model; Scherer 2003, 2013) to allow quantitative model testing.

Acknowledgments The authors would like to thank Olivier Rosset for writing the scripts for data collection (Study 2), and Tamara Ott for subject recruitment and testing (Study 2). Study 2 was supported by a Swiss National Science Foundation Grant (100014-122491) to K. R. Scherer.

References

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636.

Bänziger, T., & Scherer, K. R. (2003). A study of perceived vocal features in emotional speech. In *Voice quality: Functions, analysis and synthesis (VOQUAL'03), ISCA tutorial and research workshop, Geneva, Switzerland,* pp 169–172.

Bänziger, T. (2004). Communication vocale des émotions: Perception de l'expression vocale et attributions émotionnelles. Unpublished doctoral thesis, University of Geneva.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception. *Emotion, 12*(5), 1161–1179.

Bänziger, T., & Scherer, K. R. (2010). Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) corpus. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 271–294). Oxford, England: Oxford University Press.

Biemans, M. A. J. (2000). Gender variation in voice quality. Unpublished doctoral thesis, University of Nijmegen.

Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org/.

Davitz, J. R. (1964). *The communication of emotional meaning*. Oxford, England: McGraw Hill.

Granqvist, S. (1996). Enhancements to the visual analogue scale. *Speech, Music and Hearing: Quarterly Progress and Status Report, 4*, 61–65.

Hall, J. A., & Knapp, M. L. (2013). *Nonverbal communication*. Boston: de Gruyter Mouton.

Henrich, N., Bezard, P., Expert, R., Garnier, M., Guerin, C., Pillot, C., et al. (2008). Towards a common terminology to describe voice quality in western lyrical singing: Contribution of a multidisciplinary research group. *Journal of Interdisciplinary Music Studies, 2*(1&2), 71–93.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*(5), 770–814.

Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. In J. A. Harrigan, R. Rosenthal, & K. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 65–135). Oxford, UK: Oxford University Press.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *Journal of the Acoustical Society of America, 104*(3), 1598–1608.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge, England: Cambridge University Press.

Patel, S., & Scherer, K. R. (2013). Vocal behaviour. In J. A. Hall & M. L. Knapp (Eds.), *Handbook of nonverbal communication* (pp. 167–204). Berlin: Mouton-DeGruyter.

Robson, J., & Beck; J. M. (1999). Hearing smiles—Perceptual, acoustic and production aspects of labial spreading. In *Proceedings of the XIVth international congress of phonetic sciences*, pp. 219–222.

Rosenthal, R. (1987). *Judgment studies*. New York: Cambridge University Press.

Sangsue, J., Siegwart, H., Cosnier, J., Cornu, J., & Scherer, K. R. (1997). Développement d'un questionaire d'évaluation subjective de la qualité de la voix et de la parole, QEV. *Geneva Studies in Emotion and Communication, 11*(1). Retrieved from http://www.affective-sciences.org/system/files/1997_Sangsue_Genstudies_VoiceQuality.pdf.

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology, 8*, 467–487.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143–165.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*, 227–256.

Scherer, K. R. (2013). Emotion in action, interaction, music, and speech. In M. A. Arbib (Ed.), *Language, music, and the brain: A mysterious relationship* (pp. 107–139). Cambridge, MA: MIT Press.

Scherer, K. R., Clark-Polner, E., & Mortillaro, M. (2011). In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology, 46*(6), 401–435.

Scherer, K. R., & Ellgring, H. (2007a). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion, 7*(1), 113–130.

Scherer, K. R., & Ellgring, H. (2007b). Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion, 7*(1), 158–171.

Sundberg, J., Patel, S., Björkner, E., & Scherer, K. R. (2011). Interdependencies among voice source parameters in emotional speech. *IEEE Transactions on Affective Computing, 2*(3), 162–174.

Tartter, V. C., & Braun, D. (1994). Hearings smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America, 96*(4), 2101–2107.

Tolkmitt, F., & Scherer, K. R. (1986). Effects of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 302–313.

van Bezooijen, R. A. (1984). *Characteristics and recognizability of vocal expressions of emotion*. Doctoral dissertation. Dordrecht, The Netherlands: Foris Publications.

van Bezooijen, R. (1986). Lay ratings of long-term voice-and-speech chatacteristics. In F. Beukema & A. Hulk (Eds.), *Linguistics in the Netherlands 1986* (pp. 1–7). Dordrecht, The Netherlands: Foris Publications.